University of Calabria

Department of Mathematics and Computer Science

Master's Degree in Computer Science

Machine Learning

Machine Learning - House Prices Analysis

Daniele Avolio Alessandro Fazio Merem Hassem Indiris Michele Vitale

Lorenzo Piro

A.Y. 2022/2023

1 Introduction

To develop this **Machine Learning** project we are going to use the CRISP-DM methodology, which is a well-known and widely used methodology for data mining projects. It is an iterative process that is composed of six phases.

In particulal, the phases are:

1. **Business Understanding**: in this phase we will try to understand the problem and the objectives of the project. We will also try to understand the data that we have at our disposal and how we can use it to solve the problem.
2. **Data Understanding**: in this phase we will try to understand the data that we have at our disposal. We will try to understand the meaning of the data and how we can use it to solve the problem.
3. **Data Preparation**: in this phase we will try to prepare the data for the next phases. We will try to clean the data and to transform it in a way that will be useful for the next phases.
4. **Modeling**: in this phase we will try to build a model that will be able to solve the problem. We will try to find the best model for our problem.
5. **Evaluation**: in this phase we will try to evaluate the model that we have built. We will try to understand if the model is good enough to solve the problem.
6. **Deployment**: in this phase we will try to deploy the model that we have built. We will try to understand how we can use the model to solve the problem.

2 Business understanding

2.1 Background

Our project is based using the dataset of the Kaggle competition House Prices: Advanced Regression Techniques.

The goal of the competition is to predict the final price of each home based on 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. The dataset is composed by 1460 rows and 81 columns, where the last column is the target variable, the Sale Price.

A small list of the variables is the following:

- **LotArea**: Lot size in square feet
- **OverallQual**: Overall material and finish quality
- **OverallCond**: Overall condition rating
- **YearBuilt**: Original construction date
- **YearRemodAdd**: Remodel date
- **RoofStyle**: Type of roof
- **Exterior1st**: Exterior covering on house

- **Exterior2nd**: Exterior covering on house (if more than one material)
- **MasVnrType**: Masonry veneer type
- **MasVnrArea**: Masonry veneer area in square feet
- ...

2.2 Business objectives

In this part we are going to analyze the business objective of the project. Our goal is different from the one of the competition, in fact we are requested to **convert the Sale Price variable into 3 ranges**:

1. **LOW**: From 0 to 150000
2. **MEDIUM**: From 150000 to 300000
3. **HIGH**: From 300000 and beyond

The dataset